Analysis of time series

A *time series* is a sequence of observed values viewed as a single sample from a joint probability distribution. The central premise of a time series is that the order in which the values are observed matters. The underlying probability distribution assigns different probabilities to the same set of values if they are observed in different orders.

Let $y_1, y_2, \ldots, y_n$ denote a time series. This is a sample of size $n = 1$ from a probability distribution on the sample space $\mathcal{R}^n$. For the most part we focus here on *gridded* time series, meaning that the amount of time elapsing from y_t to y_{t+1} is the same for all values of t. Some time series are irregularly spaced or non-gridded, so we must consider the underlying sequence of time values t_i at which the data are observed.

The *mean trend* is the sequence of means of the values in the time series, i.e. the sequence $E[y_t]$ for $t = 1, 2, \ldots$. Note that this is a deterministic sequence, not a sequence of random variables. Some time series exhibit strong patterns that are best viewed as *mean trends*. For example, if y_t is the global human population in year t, say where $t = 1000, 1001, \ldots$, then y_t is increasing. We don't know for sure if this is a trend in the mean, i.e. that $E[y_t]$ is increasing, since we can only observe the history of humanity on Earth one time. But based on the nature of biological growth, it is reasonable to view the increasing values of y_t as an inevitable fact that would recur in any "replication" of the observations. That is, it is reasonable to supose that $E[y_t]$ is increasing.

Many methods of time series analysis assume that no mean trend is present, or that any mean trend present in the observed time series has been removed. If a time series has no mean trend, then $E[y_t] = c$ for all t, for some constant $c \in \mathcal{R}$. In many cases we will have $c = 0$. Such a series may still have variance and/or covariance trends, e.g. perhaps $\text{var}(y_t)$ is increasing in t, or $\text{cov}(y_t, y_{t+1})$ is increasing (or decreasing) in t.

A time series is *stationary* if for any $m > 0$, the joint probability distribution of $y_t, y_{t+1}, \ldots, y_{t+m}$ does not depend on t. For example, the probability distribution of (y_{100}, y_{101}) is the same as the probability distribution of (y_{200}, y_{201}). Note that one consequence of stationarity is that the distribution of y_t does not depend on t, and therefore in particular the variance of y_t does not depend on t, so there is a constant k such that $\text{var}[y_t] = k$ for all t. If the series is standardized then $k = 1$.

Many powerful approaches to time series analysis are based on probability models. There are many famous parametric models for time series, especially the so-called *ARIMA* models and their extensions. We will not review model-based approaches to time series analysis here. Instead we focus on approaches to time series analysis that aim to capture certain features of a time series without producing a comprehensive model for its population distribution.

Statistical analysis is *empirical* and aims to learn primarily from the data. To achieve this goal, most statistical analysis is based on exploiting some form of "replication" in the data. For example, if we wish to estimate the population mean from independent and identically distributed (IID) data, we can use the sample mean of the data as an estimate of the population mean. The replicated observations in the IID sample enable us to learn about the population mean from the sample mean. However time series are generally not IID. Fortunately, many time series exhibit a property called *mixing* that implies that forming averages over the values of a time series enables estimation of population parameters, despite the lack of IID data. Not all time series are mixing, and when a time series is not mixing most of the methods discussed here will not give meaningful results. We willnot give a formal definition of mixing here, but it is important to understand it intuitively.

Autocorrelation

If a time series is stationary, then the correlation between y_t and y_{t+1} is a constant that does not depend on t. More generally, the correlation between y_t and y_{t+d} is a constant called the *autocorrelation at lag d* that we will denote γ_d. We may also view γ_d as a function of d that is the *autocorrelation function* of the time series. This autocorrelation at lag d can be estimated by taking the sample Pearson correlation between the sequences $(y_1, \ldots, y_{n-d})$ and $(y_{1+d}, \ldots, y_n)$.

For IID data, the autocorrelation function is $(\sigma^2, 0, 0, \ldots)$, or $\gamma_j = \sigma^2 I_{j=1}$. Other commonly-encountered forms for the autocorrelation function are an exponential form $\gamma_j \propto \exp(-j/\lambda)$, or a power-law form $\gamma_j = c/(1+j)^b$.

If we consider all autocorrelations at all possible lags, we can ask whether the autocorrelations are summable, i.e. does $\sum_{j=-\infty}^{\infty} |\gamma_j|$ exist as a finite value? If the autocorrelations decay exponentially, then the autocorrelations are summable. In the power-law case, the autocorrelations are summable if and only if $b > 1$.

A time series with summable autocorrelations exhibits *short range dependence* while otherwise the series exhibits *long range dependence*. A special case of short range dependendence is known as *m-dependence*, where $\gamma_j = 0$ when $j > m$.

Robust autocorrelation

For time series with heavy tails, the conventional autocorrelation based on Pearson correlation is not very robust. A common technique that can be used in this situation is the *tau-autocorrelation*. To define this correlation measure, first consider paired data (x_i, y_i) (not time series data). Two pairs of these pairs, say (x_i, y_i) and (x_j, y_j) (where $i \neq j$) are *concordant* if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. On the other hand, the two pairs are discordant if $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$ (there are various ways of handling ties but we won't consider that here). The sample *tau-correlation* is defined to be $(a-b)/c$, where a is the number of concordant pairs, b is the number of discordant pairs, and c is the total number of pairs. There is an analogous definition for the population tau-correlation but we do not give that here.

Like the Pearson correlation, the tau-correlation lies between -1 and 1, and it population values is equal to zero when evaluating the tau-correlation between two independent random variables X and Y. As with Pearson correlation, the converse of this statement is not true – the tau-correlation can be zero even if X and Y are dependent. Positive values of the tau-correlation correspond to a type of positive association – but this is different from the positive association in Pearson correlation. The main reason to use tau-correlation arises when the data come from heavy-tailed distributions and the extreme values make it very difficult to accurately estimate the Pearson correlation.

Returning to the time series setting, we can define the tau-autocorrelation as follows. For a given lag parameter d, consider pairs of the form (x_t, x_t). Then consider the concordance of pairs of these pairs as discussed above. If the tau-autocorrelation is large (close to 1), then knowing that $x_t > x_s$ tells us that it is very likely that $x_{t+d} > x_{s+d}$.

Autoregression

One way to analyze a time series is to restructure the data into a form that can be considered using regression analysis. Most commonly, this involves partitioning the data into overlapping blocks of the form $(y_t; y_{t-1}, \ldots, y_{t-q})$. In regression terms, y_t is the response variable, and $(y_{t-1}, \ldots, y_{t-q})$ is the corresponding vector of covariates.

Autoregression can also be considered in terms of likelihoods, by factoring the joint probability distribution as follows:

$$P(y_1, \ldots, y_n) = \prod_t P(y_t | y_{t-1}, \ldots, y_1).$$

If the time series is m-dependent, we can write the above as

$$P(y_1, \ldots, y_n) = \prod_t P(y_t | y_{t-1}, \ldots, y_{t-m}).$$

If we are analyzing the data via a likelihood-based method such as maximum likelihood estimation (MLE), then the log-likelihood has the form

$$\sum_j \log P_\theta(y_t | y_{t-1}, \ldots, y_{t-m})$$

where θ is a parameter to be estimated.

Autoregression analysis can use any method for fitting regression models, for example linear modeling via ordinary least squares (OLS). Suppose we choose to fit an autoregressive model of "order m", meaning that we choose to model the conditional distribution of y_t given $y_{t-1}, y_{t-2}, \ldots$ using only the truncated history $y_{t-1}, \ldots, y_{t-m}$. If a time series is stationary and m-dependent, it makes sense to analyze it using an order m autoregression. Note that in practice we do not know if our time series is m-dependent, and if it is what is the value of m. The value of m is assessed using diagnostics and model-selection techniques. In general, if we use a given finite value of m, this does not mean that the time series must be m-dependent, but rather that we accept a small amount of bias by adopting a given finite value of m.

A basic linear autoregresive model fit using OLS uses as its dependent variable

$$y_{t+1}, y_{t+2}, \ldots, y_n$$

and the design matrix whose columns are the independent variables in the regression is

$$\begin{pmatrix} 1 & y_t & y_{t-1} & \cdots & y_{t-m+1} \\ 1 & y_{t+1} & y_t & \cdots & y_{t-m+2} \\ 1 & y_{t+2} & y_{t+1} & \cdots & y_{t-m+3} \\ & & \cdots & & \end{pmatrix}.$$

Using this response vector and design matrix, we can apply many methods for fitting regression models including OLS, PCR, dimension reduction regression, kernel methods, and many forms of regularized modeling such as the lasso and ridge regression.

Hurst parameters

A useful way to summarize the dependence structure of a time series is through the *Hurst parameter*. There are various ways to introduce the Hurst parameter and we will only consider one approach here. Recall that if we have IID data $x_1, \ldots, x_m$, the variance of the sample mean

$$\bar{x}_m = (x_1 + \cdots + x_m)/m$$

is σ^2/m. Thus, if we double the sample size to $2m$, the variance of the sample mean $(x_1 + \cdots + x_{2m})/(2m)$ is $\sigma^2/(2m)$ – the variance of the sample mean is reduced by a factor of two when we double the sample size. If we consider the log variance of the sample mean in relation to to the log sample size, we get

$$\log(\text{var}(\bar{x}_m)) = \log(\sigma^2) - \log(m).$$

Thus in log/log coordinates, the variance of the sample mean and the sample size are linearly related with a slope of -1.

It turns out that this scaling relationship between the variance of the sample mean and the sample size continues to hold as long as the dependence is "short range" as defined above. However if the dependence is long range, the variance will scale in a qualitatively different way.

If we have a sufficient amount of data, for a given block-size m we can calculate the sample means for consecutive blocks of m observations, $\bar{x}_1^m = \text{Avg}(x_1, \ldots, x_m)$, $\bar{x}_2^m = \text{Avg}(x_{m+1}, \ldots, x_{2m})$ etc., and then calculate the sample variance of these sample means:

$$v_m = \text{var}(\bar{x}_1^m, \bar{x}_2^m, \ldots).$$

Finally, we can consider the log-space relationship between $\log(m)$ and $\log(v_m)$. If $v_m = a \cdot m^b$ then $\log(v_m) = \log(a) + b\log(m)$, so b is the slope of $\log(v_m)$ on $\log(m)$. For IID and short-range dependent data, then $b = -1$ will hold. If $b > -1$ then the variances decrease slower than in the IID case, which is a logical consequence of long-range dependence. Long-range dependence implies that the time series is not mixing and does not exhibit enough independence for the sample statistics derived from different parts of the series to average to something that reflects the population structure.

The Hurst parameter is defined to be $h = 1 + b/2$, where the slope b is defined as above. When $b = -1$ (as in IID data), the Hurst parameter is $h = 1/2$. When $b > -1$, it follows that $h > 1/2$.

Periodicity

Many time series exhibit periodic behavior, meaning that for some period p, $y_{t+p} \approx y_t$ for all t. *Mean periodicity* refers to the setting where $E[y_{t+p}] = E[y_t]$ for all t.

Instead of discussing periodicity in terms of the period p, we can express it in terms of the frequency $f = 1/p$. The period is the amount of time needed to complete one cycle. The frequency is the number of cycles completed in each unit of time.

An important class of mathematical functions that are smooth and periodic are the *sinusoidal curves*. A sinusoidal curve can be written $A\cos(2\pi f t + \phi)$ where A is the amplitude, f is the frequency (number of cycles per unit time), and ϕ is the phase offset. The offset determines where in its range the time series falls when $t = 0$. For example, if $\phi = 0$ then at $t = 0$ the time series has value zero.

An important fact is that the sinusoidal curve above can be written as a linear combination of cosine and sine functions with offset $\phi = 0$, due to the identity

$$A\cos(2\pi f t + \phi) = a \cdot \cos(2\pi f t) + b \cdot \sin(2\pi f t)$$

where a, b are real scalars, with $a = A\cos(\phi)$ and $b = -A\sin(\phi)$.

It turns out that sinusoidal curves of integer-valued frequency are mutually orthogonal. This makes it convenient to use least squares regression to assess the periodicity of an observed time series.

Specifically, let $s_k(i) = \sin(2\pi k i)$ and $c_k(i) = \cos(2\pi k i)$ for $i = 1, \ldots, n$. These vectors are mutually orthogonal: if $j \neq k$, $s_j' s_k = 0$, $c_j' c_k = 0$, and $s_j' c_k = 0$ for all j, k. We can use a collection of these vectors as basis vectors for least squares regression. The fitted time series corresponding to a given basis set is

$$\hat{y} = \sum_j (s_j' y) \cdot s_j / (s_j' s_j) + (c_j' y) \cdot c_j / (c_j' c_j).$$

The *energy* (or *power*) in the observed time series y at frequency j is

$$(s_j' y)^2 / (s_j' s_j)^2 + (c_j' y)^2 / (c_j' c_j)^2$$

(there are various alternative scalings of this quantity).

A plot of power against frequency is a *periodogram*.

The above least square problem is equivalent to the *discrete Fourier transform* (DFT), and can be calculated very quickly using the fast Fourier transform (FFT). But using the FFT algorithm does not impact the interpretation of the results of this method.

If the time series is not observed at equally-spaced time points a generalization of this framework can be used. Let $t_1, \ldots, t_n$ denote the time points at which a time series $y_1, \ldots, y_n$ was observed, and define sinusoidal basis functions as $s_k(i) = \sin(2\pi k t_i)$ and $c_k(i) = \cos(2\pi k t_i)$. We can use least square regression to fit y to a set of such basis functions and produce the periodogram. In this case, the basis functions are not orthogonal and the calculation is much more expensive. There are various ways to accomplish this with the most well-known being the Lomb-Scargle periodogram.

Differencing

A simple and important technique in time series analysis is *differencing*. If our time series is $y_1, y_2, \ldots$, then the differenced time series is $y_2 - y_1, y_3 - y_2, \ldots$. We can then difference these differences, yielding second order differences $y_3 - 2y_2 + y_1, y_4 - 2y_3 + y_2$, and so on. Differencing is analogous to taking the derivative of a smooth function, and has the effect of removing longer-range trends and focusing on more local structure. It turns out that in many cases the differenced series have shorter-range dependence than the original series. At the same time differencing loses certain information about the series. In practice it may be helpful to difference one or two times and consider the structure of the original series as well as a few differenced series.